DIW Diskussionspapiere
Discussion Papers

Discussion Paper No. 79

Design-oriented Weighting of a Household Panel

by
Ulrich Rendtel*

Deutsches Institut für Wirtschaftsforschung, Berlin
German Institute for Economic Research, Berlin

Die in diesem Papier vertretenen Auffassungen liegen ausschließlich in der Verantwortung des Verfassers und nicht in der des Instituts.

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

Deutsches Institut für Wirtschaftsforschung

Discussion Paper No. 79

Design-oriented Weighting of a Household Panel

by
Ulrich Rendtel*

*) German Institute for Economic Research (DIW), Berlin

Berlin, November 1993

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14191 Berlin
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

Design–oriented Weighting of a Household Panel

Ulrich Rendtel*
German Institute for Economic Research
Berlin

November 1993

Abstract

The method of inverse sampling probabilities is adopted to calculate weights for a household panel. The method generates longitudinal as well as cross-sectional weights, which reflect the subsequent sampling stages of the panel and the different possibilities of households to enter the panel.

Key words: Weighting procedures, inverse sampling probabilities, household-panel, panel-attrition.

*The author thanks Greg Duncan and Frank v. d. Pol for helpful comments.

1 Introduction

A household panel survey differs from a series of independent cross-sectional surveys in that each wave of the panel provides the basis for the following wave of interviewing. In a household-panel some losses occur in the following wave: people die or leave the sampling-area (demographic losses) or they declare themselves unwilling to participate in the next wave (panel attrition). Also gains occur in the sample : new persons enter the panel households or panel persons leave their households and move together with other persons. Also children in panel households reach the age at which they are eligible to be interviewed. It is immediately obvious that such special characteristics of the survey have to be taken account of in the weighting procedure. I.e. the different waves of the panel must not be weighted as if they were a series of independent cross- sections. In any case this latter procedure would not help to solve the question of how longitudinal tabulations should be weighted.

This article shows how the method of inverse probalitity weighting can be adopted to a household-panel and what difficulties have to be overcome in doing this work. The approach is called "design—oriented" because some inclusion probabilities are not known a–priori by the sampling design. These are the conditional probabilities to stay in the panel in wave t given the selection in wave t-1 and the inclusion probabilities of new persons entering the the panel by moving together with people from the panel sample. This approach is also "model–based" because of the use of statistical models that estimate these components of the inclusion probabilities. Opposite to classical model–assisted approaches (cf. Särndal et al. 1992, ch. 6), which formulate models for interactions between the characteristics of the units, here the models serve to predict inclusion probabilities. In this approach the characteristics of the units are regarded as constants like in the classical randomization approach.

To estimate these models auxiliary information is used. Due to the nature of a panel survey the most relevant information is obtained from the previous panel waves and from the fieldwork of the current panel wave. Also this notion of "auxiliary information " seems unusual, since no information about population totals from sources like a micro–census is used. But it is information that is not contained in the net–sample, that is used to estimate the population totals, and this information not delivered by the sample design. So the use of the term "auxiliary information" may be justified.

The approach presented here uses extensively the sequential nature of the panel design. For each wave of the panel three models are estimated:

- a model for the recontact of households
- a model for the response of recontacted households
- a model for the inclusion probabilities of persons moving into a panel household

The probabilities derived from the above models are the modules which constitute in a simple, recursive way to compute the inclusion probabilities of the units. Due to the design—oriented approach there are different cross–sectional and longitudinal weights. For each time interval there is a different weighting scheme. This is much less than in the approach described by Lepkowski (1989), who got a different weighting scheme for every response pattern. On the other hand the number of weighting schemes is much higher than the unique weighting scheme that Pol (1993) proposed for the dutch Socio–economic Panel. Pol's approach is absolutely in the spirit of fitting the sample distribution to population marginals and ignores design aspects of a houshold panel. It also does not use any information about the non–respondents. For example it does not discriminate between losses from demographic reasons and panel attrition. The discussion at the end of this paper presents some examples which compare both strategies.

Apart from the weighting strategies that have been mentioned here, Kalton (1986) also discusses for panels the pros and cons of weighting strategies against imputation strategies. This aspect is not discussed here.

The design–oriented weighting was implemented for the the German **SO**cio-**E**conomic **P**anel (SOEP). A detailed description of this household panel, which started in 1984 on the basis of 6000 households, can be found in Hanefeld (1984) and Wagner et al. (1992). The paper also discusses the relationship of the presented weighting strategy to the weighting schemes of the **P**anel **S**tudy of **I**ncome **D**ynamics (PSID) which was started in 1968, cf. Hill (1992).

2 The computation of longitudinal weights

The aim of weighting is to estimate the number of units with a certain characteristic in the population, on the basis of the sample taken. The unknown population parameter $Y = \sum_{i=1}^{N} Y_i$ is to be estimated, where the indicator variable Y_i shows whether the i-th unit has ($Y_i = 1$) or has not ($Y_i = 0$) the characteristic of interest. Weighting by inverse sampling probabilities is based on the randomization approach. The randomization approach regards the Y_i as fixed values. Only the selection of the units is random. The random variable C_i indicates whether the unit i is sampled ($C_i = 1$) or not ($C_i = 0$). If the functional form of the estimate of Y, $\hat{Y}$, is restricted to be independent from the sample and linear in Y_i, i. e. $\hat{Y} = \sum_{i=1}^{N} \alpha_i C_i Y_i$, then unbiasedness of $\hat{Y}$ requires $\alpha_i = 1/P(C_i = 1)$. Assuming that the numbering of the units is such that the first n units are sampled one gets:

$$\hat{Y} = \sum_{i=1}^{N} \frac{1}{P(C_i = 1)} C_i Y_i = \sum_{i=1}^{n} \frac{1}{P(C_i = 1)} Y_i$$

The procedure of sampling in a panel survey can be described as a multistage process. The sample for a longitudinal interval over T panel waves is considered to be a selection process with 2T stages, which can be described as follows (the index i for the sample units has been omitted in order to simplify the notation):

Stage 1:	Design of sample (setting up the sample) $P(D = 1)$
Stage 2:	Response in the first wave $P(R_1 = 1 \mid D)$
Stage 3:	Contact successfully established in the second wave $P(K_2 = 1 \mid D, R_1)$
Stage 4:	Response given in the second wave $P(R_2 = 1 \mid D, R_1, K_2)$
.	
.	
.	
Stage 2T:	Response given in the T^{th} wave $P(R_T = 1 \mid D, R_1, K_2, \ldots, R_{(T-1)}, K_T)$

The probability of selection, $P(C = 1)$, for the whole sampling process over all subsequent stages is given by the product of the single probabilities:

$$\begin{aligned} P(C=1) &= P(D=1, R_1=1, K_2=1, \ldots, R_T=1) \\ &= P(D=1) \cdot P(R_1=1 \mid D) \\ &\quad \cdot P(K_2=1 \mid D, R_1) \\ &\quad \cdot P(R_2=1 \mid D, R_1, K_2) \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot P(R_T=1 \mid D, R_1, K_2, \ldots, R_{T-1}, K_T) \end{aligned}$$

One has to remark that temporary drop–out, i.e. units that refuse to participate in one wave and return to the sample in the next wave, is ignored here. One may integrate such participation schemes to the above situation by either ignoring the participation after temporary dropout or by treating temporary drop–out as a special kind of item-nonresponse. Alternative treatments are described in Lepkowski (1989).

The problem of weighting longitudinal sections is thus reduced to the following: Calculation of the probabilities of being sampled by design, estimation of the probabilities to respond in the first wave and estimation of the conditional probabilities of participation on the subsequent stages of the panel.

While the design probabilities are known quite well, the probabilities of response have to be estimated. Since we have only one sample at hand, we have to use a model to estimate the unknown response probabilities. The usual approach is a model, which states that the probability of response is equal for units within certain classes, defined by some adequate variables, cf. Oh/Scheuren (1983) and Särndal et al. (1992, pp. 577). Usually such within class estimates of response probabilities are very unstable due to a small number of observations within the classes. Therefore a main–effects model is used, which estimates the response probabilities from the main–effects of the variables that constitute the response classes.

There is a special difficulty in the first wave, where for non-participants only regional characteristics are known. But one may show that the fitting to marginals procedure is equivalent to the estimation of $P(R_1 = 1 \mid D = 1)$ by a main effect model in the marginals, cf. Ireland/Kullback (1968),

Oh/Scheuren (1983) and Little/Wu (1991). Thus correcting the design-weights by fitting first wave results to population marginals may be viewed as a routine to estimate $P(R_1 = 1 \mid D = 1)$.

For subsequent waves the estimation of $P(K_t = 1 \mid D, ..., R_{t-1})$ and $P(R_t = 1 \mid D, ..., K_t)$ may be performed by a logistic regression of K_t or R_t on sample information from previous waves and and fieldwork information from the present wave. It is important to note that in order to perform such an analysis one has to know the reason for drop–out. While demographic losses have to be neglected in the analysis, the losses due to non–contact or refusal are relevant for the estimation of these models. Although the fieldwork should be able to distinguish these cases, it may happen that a move out off the sampling area is not detected and recorded as panel attrition. Typically the percentage of non–recontated households is very small as compared to the number of refusals, cf. Rendtel (1990,1993a). So the potential for non–recognized moves abroad is fairly small.

There is no need to use non-sample information to estimate these conditional probabilities. It turned out that the most important reason for drop–out is a change of the interviewer, Rendtel (1990,1993a). This result is closely linked to the fact that most of the interviews in the SOEP are face-to-face interviews. But also the move of a household or the split of a household have a negative effect on participation. Although middle-class indicators were included in the logistic regression analysis of drop-out, there is no evidence of a virulent middle-class bias in participation, see Rendtel (1990,1993a).

As a consequence of this approach one gets <u>different</u> (longitudinal) weighting schemes for different longitudinal intervals. This may be viewed as a drawback since the user of weights has to decide which out of many weights is appropriate for a specific population estimate, cf. Ernst (1989). On the other hand each version of a longitudinal weight represents a different longitudinal universe [1]. So the user is forced to make precise his ideas for what population he wants to estimate totals .

The above inclusion probability is appropriate for longitudinal tabulations from the beginning of the panel until wave t. But there may be also interest to estimate totals for a longitudinal universe which starts at a later wave $1 < t' < t$. Let $C_{t',t} = 1$ denote that the unit is in the sample from wave t'

[1]The set of all persons that live in the sampling area during the longitudinal interval, see Rendtel (1993b) for different definitons of a longitudinal universe.

to wave t. $C_{t'} = 1$ indicates that the unit is in the sample at wave t'. We have to compute:

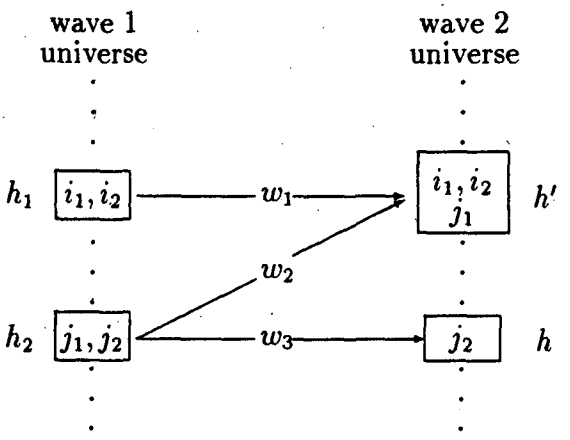
$$\begin{aligned} P(C_{t',t} = 1) &= P(C_{t'} = 1, K_{t'+1} = 1, \ldots, R_t = 1) \\ &= P(C_{t'} = 1) \cdot P(K_{t'+1} = 1 \mid C_{t'}) \\ &\quad \cdot P(R_{t'+1} = 1 \mid C_{t'}, K_{t'+1}) \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot P(R_t = 1 \mid C_{t'}, K_{t'+1}, \ldots, R_{t-1}, K_t) \end{aligned}$$

3 The computation of cross–sectional weights

The cross-sectional weight for wave t' is $1/P(C_{t'} = 1)$. Thus we see that in order to compute longitudinal weights we have to know also the cross-sectional weights. Up to now we know only how to calculate the weight of the first wave. In order to calculate weights for subsequent waves the follow-up rules of the panel come into play now. Also the fact that we deal with a household-panel reveals to be important.

The situation can be illustrated by the cross–sectional inclusion probabilities for the households in the second wave. Think of the households of the wave 1 universe and the wave 2 universe set out in two columns side-by-side (see illustration 1).

The follow-up routes
between households of
wave 1 and wave 2



There are links between some wave 1 and wave 2 households in that they have at least one person in common which would we followed if the person is in the panel.

The example given in the illlustration should make things clearer: the h_1 wave 1 household has persons i_1 and i_2 living in it, and household h_2 contains persons j_1 and j_2. Between wave 1 and wave 2 person j_1 moves together with persons i_1 and i_2 in household h', while in wave 2 person j_2 becomes the single-person household h. Therefore by the follow-up rules there are two ways in which household h' can be included in the wave 2 sample: First by household h_1 being selected in the first panel wave and subsequent follow-up via route w_1; and secondly by household h_2 being selected in the first panel wave and subsequent follow- up of individual j_1 via route w_2. Hence:

$$
\begin{aligned}
P(C_2 = 1) &= P(h' \text{ in sample wave 2 }) \\
&= P(h_1 \text{ in sample wave 1 })P(w_1 \text{ follow-up successful }) \\
&\quad + P(h_2 \text{ in sample wave 1 })P(w_2 \text{ follow-up successful }) \\
&\quad - P(h_1 \text{ and } h_2 \text{ in sample wave 1 }) \times \\
&\quad P(w_1 \text{ and } w_2 \text{ follow-up successful })
\end{aligned}
$$

Because of the generally low sampling probabilities at the start of the panel - on the average 1/5000 in the SOEP - the probabilities of jointly drawing h_1 and h_2 may be ignored, if the selection of households is not positivly correlated. This is true in general for the first wave of a panel. In later waves the selection of households is positively correlated if they stem from the same root household in wave one. There may occur the case that a child moves from the parential household and moves back some waves later. Such cases have to be differently handled; see Rendtel (1993b) for details. So the fieldwork has to be able to recognize such a return of a former household member.

In the case of a new person that enters the panel the inclusion probability of the household and its members is increased considerably. Hence the weight decreases since the increase of the sample does not reflect an increase of the population. The lower weight compensates for the over–representation of households with move–in's in the cross–sectional sample of a household panel. In the case of a person that returns to a household only a marginal increase of inclusion probability results, which may be neglected; cf. Rendtel (1993b) for details.

In a panel as a rule one observes only one household, say h_1, which contributes to h'. This happens in the situation of illustration 1 when individual $j_1 = j$ moves into household h' which is already in the panel sample. In this case the sampling probability of hosehold h_2 is unknown and may depend on household characteristics . While it could make sense to ask person j in wave 2 for the characteristics of the household h_2, in later waves the situation would be more complicated . Here one would have to ask persons for the characteristics of all households the person had lived in since the start of the panel. Therefore the SOEP does not record the characteristics of household h_2 which the individual j has left. But in the case of the SOEP household characteristics like the nationality of the head of household and the number of persons in the household considerably influence the probability of being selected in the first wave. The inclusion probabilities differ by a factor of about 10.

In this situation it is necessary to estimate the sampling probability of household h_2 in wave 1. The only information available are the individual characteristics X_j of person j. For example one may use the persons nationality which may be a good indicator for the nationality of head of the household, where the person moves from.

It is assumed here that the wave 1 inclusion probability p_j may be well fitted by a Logit-model which includes the known individual characteristics X_j of person j and also household-related characteristics H_j:

$$\ln \frac{p_j}{1-p_j} = X_j'\beta_1 + H_j'\beta_2$$

The unknown household-related component $\epsilon_j = H_j'\beta_2$ is therefore treated as an error-term in the relationship :

$$y_j = \ln \frac{p_j}{1-p_j} = X_j'\beta_1 + \epsilon_j$$

This relationship is assumed to give a good fit for all wave 1 selection probabilities. The logits y_j are known for respondents in the first wave, as well as the individual-related variables, so that β_1 can be estimated by using these variables from the wave 1 respondents. After estimation of β_1 the logit of household h_2 is estimated by $x_{j_1}'\hat{\beta}_1$. In case of the SOEP the OLS of the logits yields a fit of $R^2 \approx 0.7$. The argument to use OLS is motivated by the goal to get a good fit with small residual variance.

The procedure was displayed here for the computation of wave 2 cross-sectional weights. In later waves, say wave t, one has to regress the logits of inclusion probabilities of wave $(t-1)$ on the individual characteristics of wave $(t-1)$ sample persons. From this estimate the inclusion probabilities of new persons are estimated in wave t .

4 Analysis on household level and on individual level

Up to now we have not made explicit whether there are differences between household weights and individual weights which relate to persons. In the cross-sectional case there may be differences if the response probabilities on household- and individual-level differ, i.e. in a substantial part of the cross-section one gets answers to the household questionaire but not all household members respond to the questionaire addressed to each household member. While a uniform response pattern has been a condition to sample households at the start of the SOEP, in later waves, where this condition was dropped, it

turned out, that this pattern remains stable; i.e. either all household members participate or all household members refuse to participate, cf. Rendtel (1990,1993a). Thus with regard to cross-sectional weights household- and person-weights coincide. This coincidence of weights is different from the PSID, where new persons are treated as "non-sample-persons", and are assigned zero individual weights, cf. Hill (1992). This excludes about 20% of the sample from estimates on the personal level, c.f. Kalton (1989)[2].

If we switch to longitudinal weights we are faced with the fact that the household composition may change over time and that a sample person may live in different households in the course of the panel. Thus the question whether weights coincide for persons and households in longitudinal estimation turns out to be an ill-posed question. Although there have been attempts to define longitudinal-households and weights for these longitudinal households (see Ernst (1989)) it might be more fruitful to perform longitudinal analysis on an individual level by adding the household information to persons.

5 Discussion

Let us first look at the potential cases where the design–oriented weighting strategy will fail to give resonable estimates of the unknown probabilities of drop–out. This will occur if an important variable for drop–out is unknown for the non–respondents. Such a case will happen if the drop–out is stimulated by an event that took place since the last panel wave. For example there is some evidence in the SOEP that a divorce induces a high risk of drop out (Rendtel 1993a). This event is not recorded if both ex–spouses refuse to participate. The only indication of such an event may be reconstructed from the fieldwork of the present panel wave, where one should notice that a couple has split off into two different housholds.

For other events which are not connected with a change in the houshold composition there is even no indication of such an event. For example it

[2]One may use the PSID-weights in such a fashion that they fit into the methodological framework presented here : For cross-sectional estimates one may use the family-weights for households as well as for all household members. For longitudinal tabulations from wave 1 to wave x one may use the person-weights for wave x . Such a fashion may be justifed from a closer look at the computation of the PSID-weights, cf. Hill (1992, pp. 60-65).

might be possible — although it seems to be not very plausible in the case of the SOEP [3] — that a substantial change in the household income effects the drop–out rate. Such an influence is hard to detect by characteristics that are known for respondents and non–respondents.

One possibility to get out of this dilemma might be to compare the estimated number of events with official records. This works fine in the case of divorces. But in general official records give only information about stock of persons with a certain characteristic. Only in rare cases one gets information about the number of persons who changed their characteristics. So this strategy will work only in the case of demographic events.

Now let us see, how Pol's (1993) fitting strategy would operate in the case of divorces. If drop–out is related to a divorce, then after some waves also the stock of divorced people in the sample will differ from the official records of martial status. So the fitting strategy would adjust for the underreported number of divorced persons. But this adjustment will give higher weights to all divorced people in the sample; also to those who have been divorced before the start of the panel. This is apparently wrong since the persons, who have been divorced before the start of the panel, have no higher risk to drop out off the panel. Such a weighting may lead to wrong conclusions about the chacteristics of divorced people.

The fitting procedure does also not reflect the design–induced overrepresentation of housholds with move in's. It seems plausible that such housholds consist in their majority of younger persons. Thus a comparison with population marginals by age will reveal an overrepresentation of young persons, which will be compensated by the fitting procedure. Again all younger persons receive lower weights. But this is wrong for those who live in households without move–in's. This may lead to wrong conclusions about characteristics of young persons; for example the number of young persons which live with their parents [4]

One appealing feature of Pol's procedure is the fact that it produces — apart from a wave specific scaling factor — only one weight that is used for

[3]The empirical results suggest that drop–out in the SOEP is closely linked to a change in the interview situation which is determined by the person of the interviewer and the composition of the household, cf. Rendtel (1993a). As fas as income is concerned only an item non–response in the previous wave is relevant for drop–out. The amount of household income in wave t-1 has only a minor if any impact on the panel attrition.

[4]Since these persons will probably live in households with no move–in's.

all population estimates in the panel. But it may turn out that this is an over–simplification. The logical consequence of this procedure is that drop–out in wave $t+s$ influences population estimates in wave t. This appear to be logically inconsistent.

Finally the choice of the variables is of great concern. This is true for the design–oriented strategy and also for the fitting procedure. The answer for the design approach is: Use variable that are relevant for drop–out. The answer for the fitting procedure is different: Use good predictors for the variable of interest. The rationale behind this answer is the reduction of the variance of the population estimates. A fixed choice of age, sex, martial status and community size like in the dutch houshold panel raises doubts wether this offers a good prediction of all other variables.

Besides these doubts there exists also a second potential drawback. Suppose the following hypothetical situation: Ederly people leave the panel with increased drop–out rate.[5] This may because of illness. Now suppose that elderly persons from rich households have the tendency to be less ill because of better medical healthcare. Then :

a) The sample will under–represent elderly people.

b) The number of elderly people in rich households is in accordance with the population value.

Because of a) the fitting procedure will give higher weights to all elderly people. Thus the number of eldery persons in rich housholds is overstated. To overcome such wrong result one would need at least reliable data about the the distribution of household incomes in the population. A marginal fitting with respect to age and houshold income would however decrease the weights of all rich housholds which is incorrect again. What is necessary in this case is a population table for the joint distribution of age and household income. But usually such tables are not available.

Note that for the design–oriented approach this is no problem, since the joint occurence of age and household income is known also for the non–respondents.

[5]There is empircal evidence that this is true to some extent in the SOEP.

References

Ernst, L. (1989): *Weighting Issues for Longitudinal Household and Family Estimates.* In:Kasprzyk, D.; Duncan, G.; Kalton, G.; Sing, M. (Eds.) (1989): *Panel Surveys.* Wiley, New York ,p. 139-159.

Hanefeld, U. (1984): *The German Socio-Economic Panel.* In: American Statistical Association (Ed:): 1984 Proceedings of Social Statistics Section, p. 117-124, Washington D.C.

Hill, M. (1992): *The Panel Study of Income Dynamics; A User's Guide.* Sage Publications, Newburg Park.

Ireland, C.; Kullback, S. (1968): *Contingency tables with given marginals.* Biometrica, 55, p. 179-188.

Kalton, G. (1986): *Handling Wave Nonresponse in Panel Surveys.* Journal of Official Statistics, 2, p. 303-314.

Kalton, G. (1989): *Modelling Cosiderations: Discussion from a Survey sampling Perspective.* In:Kasprzyk, D.; Duncan, G.; Kalton, G.; Sing, M. (Eds.) (1989): *Panel Surveys.* Wiley, New York ,p. 575–586.

Oh, H.; Scheuren, F. (1983): *Weighting Adjustment for Unit Nonresponse.* In: Madow, W.; Olkin, I.; Rubin, D. (Eds.): Incomplete Data in Sample Surveys, Vol. 2, p. 143-184, Academic Press, New York.

Lepkowski, J. (1989): *Treatment of Wave Nonresponse in Panel Surveys.* In:Kasprzyk, D.; Duncan, G.; Kalton, G.; Sing, M. (Eds.) (1989): *Panel Surveys.* Wiley, New York, p. 348-374.

Little, R.; Wu, M–M. (1991):*Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ.* Journal of the American Statistical Association, 86, p. 87–95.

Pol , F. v.d. (1993): *Weighting Panel Survey Data.* Unpublished Manuscript.

Rendtel, U. (1990): *Teilnahmebereitschaft in Panelstudien: Zwischen Beinflussung, Vertrauen und Sozialer Selektion.* Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42, S. 280-299.

Rendtel, U. (1993a): *Über die Repräsentativität von Panelstichproben. Eine Analyse der feldbedingten Ausfälle im Sozi—oekonomischen Panel (SOEP).* DIW–Discussion Papers No. 70, Berlin.

Rendtel, U. (1993b): *Die Auswertung von Paneldaten unter Berücksichtigung von Panelmortalität.* Unpublished Manuscript.

Särndal, C.–E.; Swensson, B.; Wretman, J. (1992):*Model Assisted Survey Sampling.* Springer–Verlag, New York.

Wagner, G.; Burkhauser, R.; Behringer, F. (1993):*The English Language Public Use File of the German Socio–Economic Panel.* Journal of Human Resources, 28, p. 429–433.